GROUP DETECTION IN STILL IMAGES BY F-FORMATION MODELING: A COMPARATIVE STUDY

Francesco Setti[1*†], *Hayley Hung*[2*†], *Marco Cristani*[3,4*]

[1] ISTC–CNR, via alla Cascata 56/C, I-38123 Povo (Trento), Italy
[2] TU Delft, Postbus 5, 2600 AA Delft, The Nederlands
[3] Università degli Studi di Verona, Strada Le Grazie 15, I-37134 Verona, Italy
[4] Istituto Italiano di Tecnologia (IIT), via Morego 30, I-16163 Genova, Italy

ABSTRACT

Automatically detecting groups of conversing people has become a hot challenge, although a formal, widely-accepted definition of them is lacking. This gap can be filled by considering the social psychological notion of an *F-formation* as a loose geometric arrangement. In the literature, two main approaches followed this line, exploiting Hough voting [1] from one side and Graph Theory [2] on the other. This paper offers a thorough comparison of these two methods, highlighting the strengths and weaknesses of both in different real life scenarios. Our experiments demonstrate a deeper understanding of the problem by identifying the circumstances in which to adopt a particular method. Finally our study outlines what aspects of the problem are important to address for future improvements to this task.

1. INTRODUCTION

After decades studying how to automatically model single individuals (their appearance, their activities, etc.), the Computer Vision community has started to focus on how *groups* of interacting people can be modeled. In this paper, we focus on automatically identifying groups of conversing people from still images without exploiting temporal information.

There are many applications where finding such groups can be appealing such as photo tagging [3] and activity recognition [4], to name but a few. Initially, this purpose has been addressed by employing computer vision and pattern recognition tools exclusively: some of the earlier group detection methods used Voronoi diagrams with positional features [5], or modularity cut clustering [6]; subsequently, head orientation has been exploited, considering groups to contain individuals that are close and looking at each other [7, 8].

Social signal processing [9], *i.e.*, a research area that emerged at the juncture between Social Psychology and Pattern Recognition, gave new perspectives to attack the problem, inheriting social psychological notions that could inform the design of automated methods. The most important concept in this case is the notion of an *F-formation* ([10], p.209), whose original definition reads: *an F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access*;

Fig. 1. a) Example of a real F-formation with the related o-space highlighted in orange; b) frame from the *Idiap Poster Data* dataset [2]; c) frame from the *Coffee Break* dataset [1].

Roughly speaking, F-formations are spatial patterns that characterize groups of two or more people, typically gathered for conversation, to socialise, share information, and influence each other. The most important part of an F-formation is the o-space (see Fig. 1), a convex empty space surrounded by the people involved in a social interaction, in which every participant looks inward, and no external people are allowed. Associates of an F-formation are additionally defined as people who try to become involved with a conversing group but who do not succeed because they are not fully accepted by the group, or because they feel unable or unwilling in some way to converse with the others in it [10].

This paper presents a comparative study, where two of the most cited approaches for group detection by exploiting F-formations are taken into account. The first approach, called here HFF (Hough for F-Formations), consists of a Hough voting procedure which identifies F-formations by inferring their o-space center locations: this occurs by considering each person's position and their head orientation [1]. The second is called DSFF (Dominant-sets for F-Formations) [2], where people's position and body orientations are fed into a cluster-

*All authors contributed equally to this paper.

†H. Hung was partially supported by a Marie Curie Research Training Network fellowship in the project "AnaSID" (PIEF-GA-2009-255609); F. Setti was supported by the VisCoSo project grant, financed by the Autonomous Province of Trento through the "Team 2011" funding programme.

ing algorithm, whose engine depends on game theoretic dynamics and exploits the idea that F-formations are maximal cliques in edge weighted graphs (dominant sets) [11].

Both the approaches represent different interpretations of the notion of an F-formation, and this study will examine which one of them is more effective considering different scenarios: having just the body positions, adding orientations (from either head or body), and also considering different kinds of noise. The evaluation focuses on two different natural datasets: the Idiap Poster Data [2] and the Coffee Break [1] datasets; they represent quite diverse real scenarios, thus providing a solid benchmark for the two approaches.

The experiments give a clear message: with position and orientation information (even if noisy) the Hough-based procedure (HFF) provides the best results; with the body position only (*i.e.*, the most common output of body trackers nowadays) it is more convenient to employ the Dominant set-based method (DSFF). This study also helps to disect which characteristics are important for an ideal F-formation detection approach, defining clearly future perspectives for the research.

The rest of the paper is organized as follows: in Sec. 2, the HFF method is summarized, while Sec. 3 describes the DSFF method. In Sec. 4 the comparative analysis between the two approaches is presented. Finally, Sec. 5 generalizes our experimental analyses, specifying the essential characteristics a robust F-formation detector should have.

2. HOUGH FOR F-FORMATION DETECTION

In the HFF method [1], the *state* of each individual with label $i \in L$ is characterized by its floor position (x_i, y_i) and head orientation θ_i. To deal with the uncertainty over position and head orientation estimates (which presumably come from tracking and head pose estimation algorithms), a set of N samples $\{\mathbf{s}_{i,n}\}, n = 1, \ldots, N$ associated to subject i is sampled from $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_i = [x_i, y_i, \theta_i]$, and $\boldsymbol{\Sigma}$ is a diagonal matrix with trace $\sigma_x^2, \sigma_y^2, \sigma_\theta^2$. Each sample has a weight $w_{i,n} \propto \mathcal{N}(\mathbf{s}_{i,n}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, and votes for an o-space center, considering a radius R along its orientation, with an *intensity* equal to its weight. All the votes of the different subjects are stored in two accumulation spaces: an *intensity accumulation space* $\mathcal{A}_I(x, y)$ which collects the sum of the intensities of the votes for the location (x, y); and a *label accumulation space* $\mathcal{A}_L(x, y)$ that records the ID labels $\{i\}$ of the people that voted for (x, y). A final accumulation space $\tilde{\mathcal{A}}$ is built:

$$\tilde{\mathcal{A}}_I(x, y) = \text{card}(x, y) \cdot \mathcal{A}_I(x, y) \quad \forall (x, y) \in \mathcal{A}_I(x, y) \quad (1)$$

where $\text{card}(x, y)$ counts the *different* subjects that voted for x, y; such information is easily extracted from $\mathcal{A}_L(x, y)$. *Valid* groups are found by evaluating the $\tilde{\mathcal{A}}_I(x, y)$ values in descending order, and checking the o-space *emptiness condition* (which amounts to checking that people not involved in a potential F-formation instance do not lie in its o-space), iteratively, pruning away the votes of those people who have been already assigned to another legal group, until $\tilde{\mathcal{A}}_I(x, y)$ becomes empty.

In a nutshell, for the HFF method an F-formation should be instantiated by people in a good relative position (generating several votes in the $\mathcal{A}_I(x, y)$), where there is no obstacle between them (satisfying the emptiness condition).

3. DOMINANT SETS FOR F-FORMATION DETECTION

If we consider each person in the scene as a node in a graph and a non-binary distance function for the edges, an F-formation is similar to a maximal clique (or dominant set) in that graph [2]. First, an affinity matrix $\mathbf{A} = a_{ij}$ is constructed between all the nodes V. Depending if we have only position or both position and orientation information, the affinity matrix is labelled A^p or A^o, respectively. In the former case, $A^p_{ij} = -e^{d_{ij}/2\sigma^2}$, where d is the Euclidean distance between their respective positions, and σ defines the width of the Gaussian kernel centred around person i. In the latter case, a binary mask is placed over the kernel so that A^o_{ij} is only non-zero for $\pm 90°$around a person's oriented direction.

The orientation can also be estimated from the positions by assuming that people are more likely to face those that they are closer to. In such a case, a socially motivated centre of focus (SMEFO) for a person is estimated by accumulating an average position of everyone else in the scene, weighted by the corresponding A^p_{ij}; the orientation is thus described by the vector going from the person's position to their estimated centre of focus. The average weighted degree of a vertex $i \in S$ (where $S \subset V$) with respect to set S is

$$k_S(i) = \frac{1}{|S|} \sum_{j \in S} a_{ij}. \quad (2)$$

The *relative* affinity between node i and $j \notin S$ is defined by $\phi_S(i, j) = a_{ij} - k_S(i)$. The weight of each node i with respect to a set $S = R \cup \{i\}$ is defined recursively as

$$w_S(i) = \begin{cases} 1 & \text{if} |S| = 1 \\ \sum_{j \in R} \phi_R(j, i)\, w_R(j) & \text{otherwise} \end{cases}, \quad (3)$$

so S is a dominant set if $w_S(i) > 0, \forall i \in S$ and $w_{S \cup \{i\}}(i) < 0, \forall i \notin S$. To identify the dominant sets, $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$ needs to be optimized, where the non-zero elements of $\mathbf{x}$ *i.e.*, $\{x_i\}$, identifies a dominant set. A local solution of the function is found by applying the first order replicator equations taken from evolutionary game theory [11] , considering each $\{x_i\}$ in turn,

$$x_i(t+1) = x_i(t) \frac{(\mathbf{A}\mathbf{x}(t))_i}{\mathbf{x}(t)^T \mathbf{A} \mathbf{x}(t)}. \quad (4)$$

A peeling strategy identifies the dominant sets by removing each set in turn and reapplying eq. 4 on the reduced subgraph. If $w_{S \cup \{i\}}(i) > 0, \forall i \in \{V - S\}$, where V contains

all the original nodes in the graph, then the dominant set requirement is not met anymore (no more individuals can stay in a dominant set), the peeling stops, and all remaining nodes are labelled singletons.

4. EXPERIMENTS

For a thorough comparison of the two techniques, we considered different scenarios represented by two different datasets, both portraying real world scenery: the *Idiap Poster Session* dataset,[1] with ground truth annotations of position and orientation, and the *Coffee Break* dataset,[2] where real tracking and head orientation detections have been applied.

As accuracy measures, we extend the metrics proposed in [1]. The F-formation labels only considered the people who were detected by the tracker. we consider a group as correctly estimated if at least $\lceil (T \cdot |G|) \rceil$ of their members are found by the grouping method, and if no more than $1 - \lceil (T \cdot |G|) \rceil$ false subjects are identified, where $|G|$ is the cardinality of the labelled group G, and $T \in]0, 1]$ is an arbitrary threshold; in [1], $T = 2/3$, corresponds to finding at least $2/3$ of the members of a group, no more than 1/3 of false subjects. Here we also consider $T = 1$, to mean that a group is detected if all of the tracked members of an F-formation are automatically labelled. From these metrics we calculate standard precision, recall and F1 measures, on each frame, mediating them over all the frames.

4.0.1. Idiap Poster Data (IPD)
This consists of 3 hours of a real aerial video of over 50 people who met to present scientific work during a poster session. Images from the data were selected so that each one contained different F-formations. In total, 82 images were selected (to maximise on crowdedness and ambiguity) for annotation, containing ~ 1700 people. Selection was made based on leaving at least 10s between images and that no consecutively selected images contained the same formations of people. 24 annotators volunteered to label the data. The annotators were grouped into 3-person subgroups to label the same data. After being given appropriate definitions, annotators were asked to identify F-formations and their associates from static images. Asking for explicit labels for associates ensured that annotators would consciously decide how involved they thought each person was in the corresponding F-formation. Each person's position and body orientation was manually labelled and recorded as pixel values in the image plane - one pixel represented approximately 1.5cm.

4.0.2. Coffee Break (CB) *Dataset*
The dataset focuses on a coffee-break scenario of a social event, with 14 individuals at most arranged in groups of 2-3 people. The people's positions have been estimated by exploiting multi-object head tracking. Head detection has been performed afterwards [12], considering solely 4 possible orientations (Front, Back, Left, Right). The tracked positions were projected onto the ground plane before being used for our two group detectors. Even if such techniques are effective, the estimations are still noisy: for this reason this dataset represents well what a CV researcher can deal with in practice. Considering the ground-truth data, a psychologist annotated the videos indicating the groups present in the scenes, for a total of 45 frames for *Seq1* and 75 frames for *Seq2* (see Fig. 1). The annotations have been done by analyzing each frame in combination with questionnaires that the subjects filled in about the number of people they spoke with.

[1] http://www.idiap.ch/scientific-research/resources
[2] http://profs.sci.univr.it/∼cristanm/datasets.html

		Idiap Poster (IPD)			*Coffee Break* Seq1			*Coffee Break* Seq2		
Method		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
DSFF	P	**0.90**	0.81	0.85	0.53	0.54	0.53	**0.77**	0.89	0.82
	P+EO	0.88	**0.85**	**0.87**	**0.55**	**0.63**	**0.58**	0.75	0.90	0.82
HFF(-O)		0.68	0.75	0.71	0.48	0.61	0.54	0.76	**0.92**	**0.83**

Table 1. T=2/3, per frame, position only.

		Idiap Poster (IPD)			*Coffee Break* Seq1			*Coffee Break* Seq2		
Method		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
DSFF	P	**0.71**	0.65	0.68	0.29	0.32	0.30	**0.48**	0.55	0.51
	P+EO	0.70	**0.68**	**0.69**	0.32	0.40	0.35	0.48	**0.56**	**0.52**
HFF(-O)		0.47	0.52	0.49	**0.33**	**0.41**	**0.37**	0.44	0.52	0.48

Table 2. T=1, per frame, position only.

4.1. Position only

In terms of parameter settings, each method maintained its respective parameters for both data sets as the absolute values of the positions was similar. For the HFF, the following parameter values were used: $\sigma_x^2 = \sigma_y^2 = 150$, $\sigma_\theta^2 = 0.03°$, $N = 1000$ and $R = 50$, while the emptiness condition is applied over a circular area equal to 50% of the o-space. For the DSFF method, the standard deviation of the Gaussian kernel was consistently set to 40, which is twice the approximate width of a person.

First, we ran experiments assuming that only position information was available: this represents the most common scenario in a typical video-surveillance pipeline nowadays. Tables 1 and 2 show the performance when setting the threshold T to $2/3$ and 1 respectively for both methods. We consider here DSFF with the just position information (P), with the position and orientation estimated from the position (P+EO), and the HFF where a random orientation was generated by sampling uniformly around the position of each individual. We call this modified version of HFF as HFF(-O).

For both the settings, on the IPD, DSFF clearly out-performs HFF; on the CB data, the situation is not so clear; DSFF performs better on sequence 1, while HFF performs better on sequence 2 when $T = 2/3$, but the opposite is seen when using $T = 1$. This is probably due to the noisier nature of the Coffee Break dataset (many people can become occluded) and DSFF is less robust in this respect compared to HFF. The flipped ranking of the methods on the Coffee Break sequences can be explained by observing the spacing of the people in each sequence; in sequence 1 different F-formations are spaced very close together, but in sequence 2 the F-formations tend to be

	Idiap Poster (IPD)			*Coffee Break* Seq1			*Coffee Break* Seq2		
Method	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
DSFF	**0.93**	0.92	0.92	0.62	0.54	0.58	0.72	0.71	0.72
HFF	0.93	**0.96**	**0.94**	**0.67**	**0.68**	**0.67**	**0.91**	**0.91**	**0.91**

Table 3. T=2/3, per frame, position and orientation.

	Idiap Poster (IPD)			*Coffee Break* Seq1			*Coffee Break* Seq2		
Method	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
DSFF	0.81	0.81	0.81	0.39	0.35	0.37	0.39	0.39	0.39
HFF	**0.81**	**0.84**	**0.83**	**0.51**	**0.51**	**0.51**	**0.45**	**0.44**	**0.45**

Table 4. T=1, per frame, position and orientation.

spaced farther apart. The reason for the considerable drop in performance when using a more harsh threshold is probably due to the heavy occlusions, which lead to more extreme errors in the tracking, and fewer fully detected groups.

4.2. Position and orientation

Tables 3 and 4 summarize the estimation performance of both methods when position and orientation information is available. Here, HFF generally outperforms DSFF, especially in the CB sequences, while the performance on IPD is closer. This suggests that HFF is much more robust to noise; in particular, two different types of noise are present in the CB sequences: first, the tracking noise on the people's positions; second, the head orientation estimations, that are heavily discretized and so are more suited to a stochastic sampling technique. Since DSFF imposes tight constraints when detecting F-formations using the relative orientation and positioning of people in the scene, it is more sensitive to noise. This may also explain why DSFF tends to perform better on CB when using just position information, compared to both position and orientation.

4.3. Analysing the sensitivity to position-based noise

To understand the robustness of both methods to position-based noise, we performed a more systematic analysis. The IPD was preprocessed by changing the position of each person by a random value, sampled from a Gaussian distribution of increasingly higher standard deviation. Figure 2 shows the behaviour of the F_1 score as the standard deviation in pixels of the Gaussian is increased (the width of a person is roughly 20 pixels) is added to the original data. We see more clearly how HFF (blue lines) is more robust to noise when using either position only, or position and orientation information.

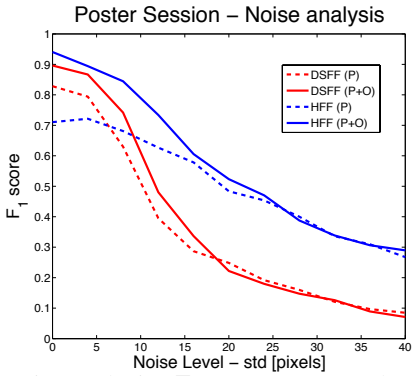

Fig. 2. Noise analysis (F_1 score) on IPD when $T = 2/3$.

5. CONCLUSION AND FUTURE WORK

In this paper, we compared the approaches of two F-formation detection methods under various conditions. HFF performed better using people's position and orientation, showing good robustness to noise. DSFF performed better when only position information was available. Our analysis highlights these generic guidelines when devising methods for identifying F-formations in still images:

- Position information is sufficient for defining groups;
- Head or body orientation enables improvements in detection performance, especially in very crowded scenes;
- Since position and orientation information may be affected by noise, a robust management of the uncertainty should be considered in the group estimation process.

We plan to follow these guidelines by joining the strengths of both DSFF and HFF for more robust F-formation estimation, since both methods contain key qualities for group detection.

6. REFERENCES

[1] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of F-formations," in *BMVC*, 2011.

[2] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in *ICMI*, 2011.

[3] M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "Here's looking at you, kid. Detecting people looking at each other in videos," in *Proc. of BMVC*, 2011.

[4] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proc. ACM Int. Workshop on Multimodal Pervasive Video Analysis*. 2010, pp. 37–42, ACM.

[5] J. C. S. Jacques, A. Braun, J. Soldera, S. R. Musse, and C. R. Jung, "Understanding people motion in video sequences using Voronoi diagrams: Detecting and classifying groups," *Pattern Analysis and Applications*, vol. 10, no. 4, pp. 321–332, 2007.

[6] T. Yu, S. Lim, K. A. Patwardhan, and N. Krahnstoever, "Monitoring, Recognizing and Discovering Social Networks," in *Computer Vision and Pattern Recognition*, 2009.

[7] N. M. Robertson and I. D. Reid, "Automatic Reasoning about Causal Events in Surveillance Video," *EURASIP J. Image and Video Processing*, 2011.

[8] Khai N. Tran, Apurva Bedagkar-Gala, Ioannis A. Kakadiaris, and Shishir K. Shah, " Social Cues in Group Formation and Local Interactions for Collective Activity Analysis," in *VISAPP*, 2013.

[9] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *IVC*, 2009.

[10] Adam Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*, Cambridge University Press, 1990.

[11] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *PAMI*, 2007.

[12] D. Tosato, M. Spera, M. Cristani, and V. Murino, "Characterizing Humans on Riemannian Manifolds," *PAMI*, 2012.